Лабораторная работа №3
«Методы дискриминантного анализа»

Цель работы: получить практические навыки работы с методом дискриминантного анализа и визуализацией данных на практических примерах с использованием языка программирования python.

Теоретический материал

Дискриминантный анализ, как раздел многомерного статистического анализа, при «классификации с учителем» включает в себя статистические методы классификации многомерных наблюдений. Например, для оценки финансового состояния своих клиентов при выдаче им кредита банк классифицирует их по надежности на несколько категорий по ряду признаков. В случае, когда следует отнести клиента к той или иной категории используют процедуры дискриминантного анализа.

Все процедуры дискриминантного анализа можно разбить на две группы и рассматривать как совершенно самостоятельные методы. Первая группа процедур позволяет интерпретировать различия между существующими классами, вторая – производить классификацию новых объектов в тех случаях, когда неизвестно заранее, к какому из существующих классов они принадлежат.

Пусть имеется множество единиц наблюдения. Каждая единица наблюдения характеризуется несколькими признаками: x_{ij} – значение j-й переменной i-го объекта (i = 1, ..., n; j = 1, …, p). Предположим, что все множество объектов разбито на несколько подмножеств (два и более). Из каждого подмножества взята выборка объемом n_k, где k – номер подмножества (класса), k = 1, …, q.

Признаки, которые используются для того, чтобы отличать один классов (подмножество) от другого, называются дискриминантными переменными. Число объектов наблюдения должно превышать число дискриминантных переменных: p < n. Дискриминантные переменные должны быть линейно независимыми. Основной предпосылкой дискриминантного анализа является нормальность закона распределения многомерной величины. Это означает, что каждая из дискриминантных переменных внутри каждого из рассматриваемых классов должна быть подчинена нормальному закону распределения.

Основная идея дискриминантного анализа заключается в том, чтобы определить, отличаются ли совокупности по среднему какой-либо переменной (или линейной комбинации переменных), и затем использовать эту переменную, чтобы предсказать для новых членов их принадлежность к той или иной группе. Канонической дискриминантной функцией называется линейная функция:

$$\boldsymbol{d_{km} = \beta_0 + \beta_1 * x_{1km} + \cdots + \beta_p * x_{pkm},}$$

где:

$\boldsymbol{d_{km}}$ – значение канонической дискриминантной функции для m-го объекта в группе k (m = 1, …, n, k = 1, …, g)

$\boldsymbol{x_{pkm}}$ – значение дискриминантной переменной X_i для m-го объекта в группе k

$\boldsymbol{\beta_0, \dots, \beta_p}$ – коэффициенты дискриминантной функции.

С геометрической точки зрения дискриминантные функции определяют гиперповерхности в p-мерном пространстве. В частном случае при p = 2 она является прямой, а при p = 3 – плоскостью.

Коэффициенты β_i первой канонической дискриминантной функции выбираются таким образом, чтобы центроиды (средние значения) различных групп как можно больше отличались друг от друга. Коэффициенты второй группы выбираются также, но при этом налагается дополнительное условие, чтобы значения второй функции были некоррелированы со значениями первой. Аналогично определяются и другие функции. Отсюда следует, что любая каноническая дискриминантная функция d имеет нулевую внутригрупповую корреляцию с d_1, d_2, …, d_{g-1}. Если число групп равно g, то число канонических дискриминантных функций будет на единицу меньше числа групп. Однако по многим причинам практического характера полезно иметь одну, две или же три дискриминантных функций. Тогда графическое изображение объектов будет представлено в одно-, двух- и трехмерных пространствах. Такое представление особенно полезно в случае, когда число дискриминантных переменных p велико по сравнению с числом групп g.

Ход работы

1. Прочитать теоретическую часть по методам дискриминантного анализа.
2. Описать структуру исходных данных для своего набора:
    a. общие характеристики массива данных: предметная область, количество записей
    b. входные параметры: названия и типы
    c. выходной класс: название и значения
3. Осуществить визуализацию двух любых признаков и посчитать коэффициент корреляции между ними
4. Выполнить разбиение классов набора данных с помощью LDA (LinearDiscriminantAnalysis). Осуществить визуализацию разбиения
5. Осуществить классификацию с помощью методов LDA и QDA (LinearDiscriminantAnalysis и QuadraticDiscriminantAnalysis). Сравнить полученные результаты